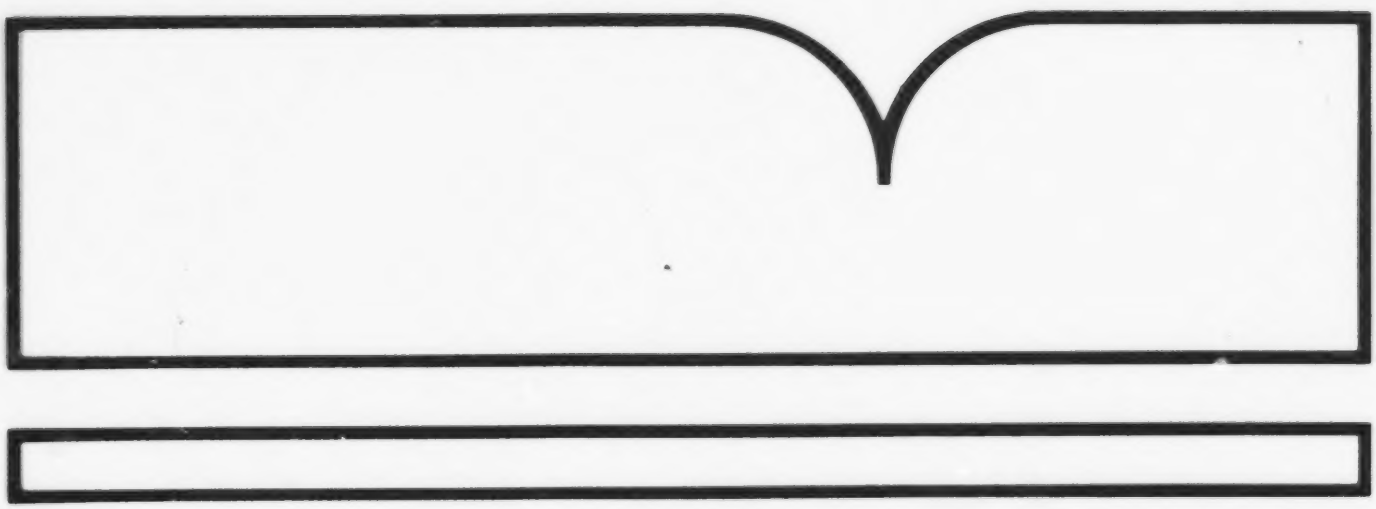
PB83-209817

Overview of In vitro Teratogenicity Testing: Aspects of Validation and Application to Screening

(U.S.) Health Effects Research Lab.
Research Triangle Park, NC

1982



U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before comple

1. REPORT NO. EPA-600/J-82-315	2. JOURNAL ARTICLE	3. PB83-209817
4. TITLE AND SUBTITLE: Overview of In Vitro Teratogenicity Testing: Aspects of Validation & Application to Screening		5. REPORT DATE: 1982 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S): G. L. Kimmel, M. Kate Smith, I.D. M. Kochhar and R. M. Pratt		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Kimmel - NCTR; Smith - U.S. EPA, Cinti., OH; Kochhar - Jefferson Medical College; Prattt - NEIHS		10. PROGRAM ELEMENT NO. CCSLIA 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS: Health Effects Research Laborattory, Office of Research & Developmennt, U.S. EPA, Research Triangle Park, NC 277711		13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA-600/11

15. SUPPLEMENTARY NOTES

Teratogenesis, Carcinogenesis, and Mutagenesis 2:221-229 (1982)

16. ABSTRACT

Overview of In Vitro Teratogenicity Testing: Aspects of Validation and Application to Screening.

A summary of response to questions which arose as a result of a workshop on Teratogenicity Testing and discussion relative to those questions.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT: Release to Public	19. SECURITY CLASS (This Report) unclassified	21. NO. OF PAGES 10
	20. SECURITY CLASS (This page) unclassified	22. PRICE

EPA Form 2220-1 (9-73)

Teratogenesis, Carcinogenesis, and Mutagenesis 2:221–229 (1982)

EPA-600/J-82-315
JOURNAL ARTICLE

Overview of In Vitro Teratogenicity Testing: Aspects of Validation and Application to Screening

G.L. Kimmel, K. Smith, D.M. Kochhar, and R.M. Pratt

Interest in the application of in vitro systems to teratogenicity testing was evident as early as 1975, when the International Conference on Tests of Teratogenicity In Vitro was held to bring together investigators with experience in in vitro systems as applied to the developing organism.* Presentations at this conference focused on the application of in vitro systems to basic developmental questions, and while they established a good base of information about the characteristics of these systems, the question of their applicability to compound screening did not receive significant attention. More recently, it has become evident that the current in vivo teratology test procedures are overburdened by the large number of compounds that enter the environment annually. A number of investigators have therefore begun establishing systems that may serve as teratogenicity screens. Unfortunately, the methods for validating such systems are as yet undefined. Thus, the need to establish a stronger base of information from which development and validation could take place was recognized and provided the framework for the workshop.

As indicated in the Foreword of this issue of *Teratogenesis, Carcinogenesis, and Mutagenesis*, specific questions served as a focus for the workshop, and the overview summarizes the discussion relative to these questions. The questions were interrelated; consequently, the discussion often addressed aspects of several questions. Therefore, while the overview will generally follow the order of the questions, it will not be limited to concise responses, but rather will attempt to indicate areas of consensus and areas of disagreement. It should also be noted that although the articles that follow the overview describe systems that may prove useful as tests of teratogenic potential, these do not include all of the systems that should be considered as in vitro approaches to hazard

*Proceedings on the International Conference on Tests of Teratogenicity In Vitro, eds. E.D. Ebert and M. Marois, North-Holland Publishing Company, New York, 1976.

1982 Alan R. Liss, Inc.

evaluation. The systems that were chosen for presentation were intended to provide a representative range of in vitro approaches and a basis for discussion of the questions that were being posed.

QUESTION ONE

> What can be considered an acceptable measure of teratogenic potential, ie, quantifiable endpoint, for in vitro test systems? How closely should these endpoints be related to embryogenic events in the "intact mammalian system"? If a test with an apparent "developmentally irrelevant" endpoint proves to respond well to validation criteria, should it be considered for inclusion in a test battery for regulatory purposes?

The wide variety of in vitro systems used to study normal and abnormal development has a correspondingly large number of endpoints that can be used to evaluate adverse developmental effects. These range from such endpoints as somite number and neural tube closure in the whole embryo to characteristics of growth and attachment in cell culture. Consequently, to define particular endpoints as "acceptable" is a difficult task, given the current state of our knowledge. The workshop participants refrained from such a definition, and the group that had primary responsibility for this question developed a general statement that could be applied to all in vitro test systems.

> An acceptable measure of teratogenic potential in vitro is an endpoint that would predict findings of generally accepted evaluations of teratogenic potential in vivo. The endpoints may not necessarily be related to embryonic events, although in most cases an ontogenetic system would be expected. The endpoints must be reproducible, capable of statistical analysis and capable of detecting a dose-response relationship. They should eventually predict human teratogenicity.

There was general agreement with this statement, although several aspects were considered in greater detail. For instance, further definition was made of what constituted a "generally accepted evaluation of teratogenic potential in vivo." The basis for comparison should be the in vivo models most widely used and for which the greatest amount of data was available, ie, the mouse, rat, and rabbit. Nevertheless, where possible, data collected in humans and subhuman primates should be referred to and used as may be appropriate.

The importance of quantifiable endpoints to both assay validation and the ranking of an agent's teratogenic potential was recognized. Generally, endpoints expressed as percentages, ratios, or proportions were preferred to provide the type of information necessary for these types of determinations. At present, any endpoint that could demonstrate a dose-response relationship and be statistically analyzed would be acceptable. There was some disagreement on whether an endpoint could take the form of a binomial, ie, yes-or-no designation. Some participants felt that while an endpoint that could show a dose-response relationship was clearly preferable, it was premature at this stage of assay development to exclude endpoints that take the binomial form. Other participants, however, questioned the amount of information that could be obtained from binomial endpoints and its usefulness in validation or testing. They felt that a numerical form was important in order to permit the comparison of data from in vitro systems with data based on in vivo mammalian studies.

When considering the "developmental relevance" of endpoints and their relation to embryogenic/ontogenetic events, the choice of an acceptable endpoint was further

confounded. It appeared at first that the endpoints measured in a number of the systems bear little direct relationship to teratogenic potential (ie, production of malformations, functional deficits, etc). A cell culture system, for example, is far removed from measuring a malformation; rather, it measures an event that may or may not be relevant to the production of malformations. The view was widely expressed that "intuitively" a useful system would be expected to employ an ontogenetic change as an endpoint. However, after considerable discussion two things became apparent. First, virtually any normal cellular or subcellular event could be considered an integral part of development. Second, on the basis of our present knowledge of teratogenic mechanisms, there is no justification for demanding that a test be directly and closely related to the developing embryo/fetus in vivo. The advantage of such a relationship was recognized, but it was the clear consensus that if tests using a "developmentally irrelevant" endpoint could predict the findings of generally accepted safety evaluations of teratogenicity, then the endpoint was acceptable.

It can be further noted that the requirement of relevance to ontogeny may also be influenced by how a test is to be used. For example, if a single test is to be used as the sole indicator of teratogenic potential, the need for developmental relevance may be great. However, if a series or battery of in vitro tests were used to supply a composite answer, then such tests may complement each other in the type or range of event measured.

QUESTION TWO

> To be accepted as part of a screening process, in vitro systems will have to be validated for their ability to predict the teratogenic potential of an agent. What are the necessary criteria for validation? How will the choice of compound used in the validation be made, by structure, biologic activity, environmental use, etc? How many compounds should be used for validation and what percentage of false negatives and false positives will be acceptable? On what basis will dosing be defined and how will it relate to the in vivo situation?

It was clear from discussion of the previous question that well-defined endpoints are required for an in vitro system to be useful in the screening of environmental agents, and that these endpoints should predict laboratory findings of in vivo teratology and eventually the potential for human teratogenicity. This same predictive requirement is central to the criteria for test validation, and it became a focus of discussion for the second question.

It was the consensus of the workshop that, to carry out validation in a controlled fashion and to facilitate interlaboratory comparisons, a list of agents that could be used for validation should be developed by a panel of investigators experienced in teratology with input from individuals experienced in pharmacokinetics, metabolism, and the characteristics of chemical classes. This list would then be made available to any laboratory with interest in validating a particular test system. The agents should be chosen on the basis of their in vivo teratogenic activity and care should be taken to assure the reliability of the in vivo studies reviewed. The agents should range from established teratogens to agents showing only marginal teratogenicity. The selection of negative controls was clearly a problem for most investigators and was viewed as an area in which the suggested panel could provide considerable guidance. To the extent possible, the agents should be ranked relatively to their teratogenicity. Agents repre-

sentative of the various physical states, ie, gas, solid, and liquid, should be considered. The choice should not be made on the basis of molecular structure or structure-activity relationships, since this has not been a reliable predictor of teratogenic activity in the past. However, it was noted that chemical structure should be considered, to assure that the list included compounds from a variety of chemical classes. Choice based on environmental use was also discouraged, although it was decided that where information is available, it may be noted.

A specific number of agents was not defined, and will depend on several factors, including achievement of a diversity of physical, chemical, and biological characteristics. It was considered that as many as 50 compounds, covering the range of teratogenicity, might be chosen. Several workgroups suggested specific agents or groups of agents that could be considered for inclusion, and these were forwarded for the panel. It was noted that at a minimum all undisputed human teratogens should be included. Once a final list has been established, then it was suggested that a compound repository be developed from which all investigators could obtain agents for assay validation. This would permit the determination of an agent's purity and its solubility and stability in the vehicle of administration, providing a strong base for interlaboratory comparison of test systems.

The amount of latitude that an investigator should have in choosing agents from the list for validation of a particular system was discussed. The point was made that all tests should be validated on a uniform basis, thus requiring that *all* agents on the list be tested within a given system. This would provide a more reasonable estimate of the predictive nature of a specific test, and indicate the degree to which that test may be valid for most substances but invalid in its testing of a particular class of compounds. However, while this may be ideal, some felt that from a practical standpoint this may not be possible, owing to the considerable time and resource requirements of such an approach. In other words, while testing the entire list of established agents is preferable, the validity of a particular system might still be considered before all agents had been tested.

The acceptable level of false-positives and false-negatives (ie, agents whose teratogenic activity in vivo is not accurately assessed in vitro) was then addressed. It was generally agreed that a specific number could not be assigned to the permissible level of either false-positives or false-negatives, although it was understood that an overabundance of either may bring a test into question. The specific number would depend on several factors, including the nature of the assay being considered, its use as a single test or as part of a test battery, and its placement in the overall toxicity assessment process. For example, the chick embryo is known to be highly sensitive to a number of agents, thus potentially producing a large number of false-positives. If this assay were used as the only predictor of the teratogenic potential of an agent, this large number of false-positives would reduce its utility. However, in combination with other assays, the highly sensitive nature of the chick embryo system may prove to be of great benefit to the overall assessment process. Likewise, it is generally agreed that false-negatives can pose one of the most serious problems in screening systems, since teratogenic agents may go undetected. A test producing a large number of false-negatives might be of little utility unless it was coupled with complementary testing. A compound tested in such a system would not be judged harmless on the basis solely of the lack of a positive response. Finally, the "acceptable" level of false-positives and false-negatives is obviously related to the accuracy of in vivo test procedures currently in use. A test or test battery

may be "acceptable" simply because its accuracy is an improvement over generally accepted evaluations of teratogenicity.

The correct basis for dose selection in vitro and its relationship to in vivo exposure is dependent on the specific system being considered. It was thought that, once data on dose-response relationships are collected, defining a "most appropriate" basis for dosing might be possible. However, some of the participants felt that determination of such a basis might still be difficult. For example, how does one relate dosing on the basis of milligrams/kilogram in the intact mammal with micrograms/milliliter in a cell culture system? It was reiterated that any dosing design that would "predict the findings of generally accepted safety evaluations of teratogenicity" would be relevant. This issue may be better resolved when comparisons can be made of dose-response data for the same compound using different methods.

QUESTION THREE

> How could in vitro systems be applied to screening previously untested individual compounds or mixtures of compounds with unknown teratogenic potential? Consideration should be given to the specific experimental design, to the choice and administration of doses, and to the distinction of general toxicity from specific endpoints of developmental toxicity. Could these systems be used to prioritize the further testing of unknown compounds?

Once validation has been accomplished, systems that are useful as predictors of teratogenic potential can be applied to the screening of agents that have previously been untested for teratogenicity. With this in mind, the discussion concentrated on the manner in which validated in vitro tests may be applied and on their placement in the toxicity assessment process. Given the current state of development of in vitro test systems, the consensus was that all tests should be considered preliminary indicators of teratogenic potential. No single in vitro system or combination of systems was considered capable of reducing the need for continued in vivo testing. Rather, in vitro assays were seen as aiding in a more effective testing program by indicating the priority for further, more extensive testing.

The use of a single test versus a battery of tests was raised early in the discussion, and the participants felt that currently no single test could be considered likely to serve as a screen but that a battery of tests would be required. This battery should include tests with endpoints reflecting most of the developmental phenomena that if perturbed could lead to a teratogenic event. The battery would probably be composed of two or more separate and diverse biologic systems that collectively would display several embryologic processes, eg, cell proliferation, susceptibility to cell death, tissue differentiation, cell adhesion, organogenesis, etc. For instance, a battery may include a rapid and simple cell culture as well as an assay that measured aspects of progressive development. The first would provide a rapid measure of biologic activity of the compound as it relates to such cellular parameters as viability and proliferation. The latter would provide a more detailed evaluation of ontogenetic events that are important to normal development. It was noted that the tests that make up this battery should be considered on a basis of their being complementary and supplementary to each other. Thus, a battery of tests designed to be complementary would ensure that false-negatives in one assay would exhibit a positive response in at least one other assay in the battery. In contrast, assays within a battery providing the same determination for a particular test sub-

stance would be considered supplementary. While this is not considered as important as complementarity of the assays, it should be considered a significant advantage and noted in the development of any battery.

In addition to the criteria already mentioned, it was further noted that the design of an in vitro screening battery should be simple, allowing for modifications as knowledge in the field increases. Simplicity is an important aspect for intralaboratory reproducibility and strengthens an assay's application and comparison under interlaboratory conditions. The assays should be of relatively short duration and the battery should be applicable to all agents. The dosing or exposure procedure should be uniformly applied to all agents, varying only according to the solubility and other physical characteristics of the agent. The dose range should be from a no-effect level to a level providing an obvious effect. This will assist in establishing a dose-response relationship and will allow for the ranking of various agents according to their teratogenic potential.

As mentioned above, in vitro systems were seen not as replacements for current in vivo testing but rather as a means of prioritizing chemicals in a more effective testing program. In recognition of this, aspects of the design and placement of a test battery in the overall test scheme were discussed, and two workgroups provided examples of possible approaches to teratogenesis testing. These examples were offered as a base for further comment, criticism, and discussion and they served that function very well. The first example detailed an overall approach to teratogenesis testing in which agents of unknown teratogenic potential would first be assayed in a battery of in vitro tests. Those agents that showed little or no activity would be assigned a low priority for further testing unless it was contraindicated, by additional considerations (eg, higher than expected industrial or environmental exposure). Agents that demonstrated a positive activity in vitro would then be considered in relation to their expected level of exposure in the environment and any test data on other types of toxicity. In this way, some of the positive agents may be given a lower priority for further testing, either because the level of exposure in the environment is too low to cause immediate concern or because other types of toxicity (eg, carcinogenicity) might preempt the use of data on teratology in the overall risk assessment process. Those positive agents that remained would then be given a high priority for further testing in vivo. It was noted that at this point full-scale, in vitro testing might be preceded by short-term in vivo assays (eg, Chernoff and Kavlock, *Teratology*, 21:33A–34A, 1980), which could further facilitate prioritization.

Considerable discussion focused on the suggestion that a compound that tested positive in vitro could be given a low priority for further testing, on the basis of strictly its suspected carcinogenicity or mutagenicity. This seemed to suggest that testing for teratogenicity was important only for those agents that demonstrate no carcinogenic or mutagenic activity. Several participants indicated that this may indeed reflect reality in the use of toxicity data in risk assessment, regardless of whether it was the most appropriate scientifically. In contrast, other participants noted that while this may have been the approach to previous decisions in particular cases, in general individuals involved in risk assessment find it important to develop a toxicologic profile that includes as many aspects of toxicity as possible. Thus, the overall hazard of exposure can be assessed. It was also pointed out that decisions on priority testing based on such criteria as low levels of environmental exposure or potential for accidental exposure might be made prior to any testing, including in vitro; and it was agreed that such considerations would be important at all points in the testing process.

Another test design put forward was a tier approach, and one workgroup used this as a base for much of their discussion. In this approach, the initial testing would be carried out in cell cultures, organ cultures, or submammalian organisms. At this level, the test system does not have to show developmental relevance (ie, progressive development). Those agents that tested positive would be taken to a second tier of in vitro tests that would be considered developmentally relevant, eg, organ culture and whole-embryo culture. Following testing at this level, all suspect agents would then be tested in vivo.

During the discussion of this approach, concern was raised regarding the occurrence of false-negatives in the first tier. If only those agents showing a positive result at tier I are carried on to tier II, then agents that were false-negatives in tier I would be excluded from further testing and their teratogenic potential missed. The discussion group had considered this possibility but felt that an agent would be unlikely to show a false-negative if a sufficient number of assays were included in tier I. In this manner, the level of false-negatives in one specific assay may be high, but the cumulative total of false-negatives in a tier I battery would be quite low. The assays that would be included at this level should be highly sensitive to perturbation, reducing the chance for false-negative results. Some participants expressed concern that tiers within in vitro testing should be considered at all, given the current state of our knowledge. They felt that tests with different endpoints, sensitivity, and developmental relevance could be included in a test scheme without further definition of tiers and that a tiered approach would not necessarily be an advantage to the overall testing process. One advantage might be that the simpler design and resource requirements of the tier I assays may make this approach less costly than other test schemes. It was also noted that presently identification, confirmation, and assessment of a particular teratogenic hazard must all be accomplished by the in vivo test. If a group of assays could be developed that would identify a potential hazard in tier I and confirm that potential hazard in tier II, this overall scheme of tiered testing might have a certain advantage.

In summary, discussion of these two proposals did not lead to a consensus on the most appropriate test scheme. However, it did highlight two points that were raised throughout the workshop discussion. First, it is important to develop theoretical models of validation and application. Such theoretical examples provide a base, as did these examples, for critically evaluating various aspects of the proposed models. Second, given the current state of in vitro test development and validation, there must be latitude for changes that will inevitably come with further research and application.

QUESTION FOUR

Is there a requirement that in vitro tests mimic the metabolic capacity of the in vivo system?

The need for representative metabolism in an in vitro test system was generally viewed as a requirement. It was recognized that in a battery or tiered approach not every test may need to meet this requirement, since the time and resources required might be prohibitive. Nevertheless, the overall evaluation of the screen must account for aspects of metabolism if this evaluation is to have any utility in prioritization for further testing.

The workgroup that had primary responsibility for this question presented conditions under which metabolic activating systems should be used. It was suggested that all test compounds should be tested both with and without a metabolic activating system and that in those test systems where such activity was not endogenous, an exogenous source should be added. A suggested modification of this approach was to include representative metabolic capabilities only in the case where no effect was seen at high concentrations or at the limit of solubility. The group recommended that one standardized metabolic system be used exclusively, ie, the S-9 fraction obtained from male rats of the Sprague-Dawley strain after induction by Aroclor 1254. On this point, there was considerable discussion and it was pointed out that in some cases such metabolic activating systems are incompatible with the viability of the assay or with the accurate evaluation of teratogenicity. It was also noted that while structure-activity relationships have not proven successful in determining teratogenic activity, consideration of the structure of a compound can be helpful in determining the type of metabolic activating system that may be required. For example, while an oxidative environment may very well be the most appropriate metabolic system for certain compounds, other compounds may require reductive reactions. It is also possible that activation is mediated through a multistep reaction that requires steps either before or after the oxidative activity. In both cases, the teratogenic potential would be missed by addition of only the standard S-9 fraction. It was also pointed out that cellular subfractions may overrepresent the production of reactive intermediates while underrepresenting reactions that tend to detoxify the reactive intermediates. In vitro systems that maintain their endogenous capacity for metabolic activation and detoxification are important for addressing some of these concerns. For example, some systems are amenable to the use of serum from animals that have been exposed in vivo to particular agents, allowing endogenous metabolism to be effected prior to exposure of the in vitro subject.

Finally, while a consensus was reached that knowledge of metabolism and metabolic activation of agents to proximate teratogens was important, it was noted that this particular area of systems development is probably one of the least defined and understood, at least when relating these systems to the screening of potential teratogens. This is true not only of the specific events in metabolism, but also of the the ontogeny of these metabolic systems and their variability between species. Thus, it is apparent that this area stands out as one that will require increased attention in the future.

In summary, the workshop was organized to address several basic questions relative to the development and validation of in vitro test systems as screens of teratogenicity. While more questions were raised than answered, the three days of discussion helped focus the current state of the art and provide direction for the future development, validation, and application of these systems. The workshop makes us much more aware of the variety of in vitro systems that exist, and the considerations that are applicable to all, as well as those that are specific to only a few. A clear definition of an acceptable in vitro endpoint has been developed and will serve to guide the future validation process of all in vitro assays. The need for an established list of compounds that can be used in the validation of all systems was also recognized and the organization of a panel to develop such a list has been initiated. Much of the discussion did not lead to concise answers to the questions. However, it was the consensus that many of the points, including developmental relevance, quantitative endpoints, metabolism, and the total process of validation and application, were now more clearly defined in a manner that would permit their consideration in the future. It was further noted that

relative to the application of these systems as screens, many of these points were being addressed for the first time. Ultimately, it should be possible to develop in vitro assays that can serve an important function in an effective testing program. At a minimum, our knowledge of developmental processes, including metabolism and mechanisms of action, should improve considerably and place us in a better position to determine the teratogenic potential of an agent and the possible human risk. The requirements for such an approach, ranging from research support to the development of more appropriate methods, should be recognized, and the process of test validation and application should now receive increased attention.